An Effective Algorithm for Semantic Similarity Metric of Word Pairs*

Lingling Meng[1], Runqing Huang[2] and Junzhong Gu[3]

[1]Computer Science and Technology Department, Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China
[2]Shanghai Municipal People's Government, Shanghai, 200003, China
[3]Computer Science and Technology Department,
East China Normal University, Shanghai, 200062, China
llmeng@deit.ecnu.edu.cn, runqinghuang@gmail.com, jzgu@ica.stc.sh.cn

Abstract

Semantic similarity is fundamental operation in the field of computational lexical semantics, artificial intelligence and cognitive science. Accurate measurement of semantic similarity between words is crucial. The paper presents an effective algorithm for semantic similarity metric of word pairs. Different from previous work, in the new algorithm not only path length, but also IC values have been taken into account. We evaluate our model on the data set of Rubenstein and Goodenough, which is traditional and widely used. Coefficients of correlation between human ratings of similarity based on seven algorithms are calculated. Experiments show that the coefficient of our proposed algorithm with human judgment is 0.8820, which demonstrate that our new algorithm significantly outperformed others.

Keywords: *semantic similarity, path based, information content based, WordNet*

1. Introduction

Semantic similarity has attracted great concern in artificial intelligence and cognitive science for many years. It can be dated back to Quillian [1] and the spreading activation algorithm. Nowadays, semantic similarity has been successfully applied in word sense disambiguation [2], information extraction [3, 4], semantic annotation and summarization [5, 6], question answering [7], recommender system [8], text segmentation [9] and so on. It shows its talents and makes these applications more intelligent. Therefore it is necessary to design accurate methods for improving the performance of the bulk of applications relying on it. This paper presents an effective algorithm for semantic similarity metric of word pairs. Both path length between two concepts and their information contents have been taken into account. Experiments show that our new algorithm significantly outperformed related work.

The rest of this paper is as follows: in Section 2 related work of measuring semantic similarity between word pairs are discussed. A novel semantic similarity algorithm is presented in Section 3. Section 4 shows the evaluation of the new method, including the experiments, data analyzing, and the advantage. Conclusion and future Work is described in Section 5.

* The work in the paper was supported by Shanghai Scientific Development Foundation (Grant No.11530700300)

2. Semantic Similarity Algorithm

It is a challenge to accurately understand concepts expressed in natural language. Generally speaking it is usual to be decompressed into comparing semantic similarity between concepts. How to measure concepts similarity? Many algorithms have been proposed. In terms of how to utilize knowledge base, all the algorithms can be classified into two categories: knowledge-poor methods and knowledge-rich methods [10]. Knowledge-rich methods require semantic networks or a semantically tagged corpus to define the concepts relations, such as WordNet, Roger, Longman and so on. Recent years the methods based on WordNet have drawn great concern. The algorithms discussed in the paper are all based on WordNet.

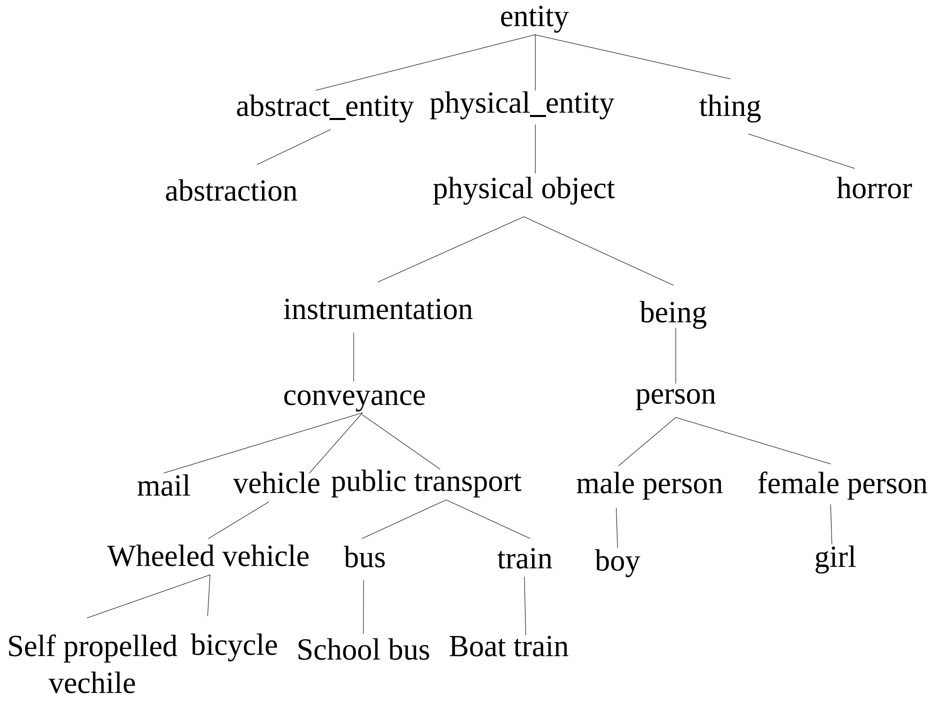

Figure 1. A Fragment of is-a Relation in WordNet

WordNet is the product of a research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English [11]. Now it has become a valuable resource and played an important role in human language technology. WordNet focuses on the word meanings instead of word forms. In WordNet nouns, verbs, adjectives, and adverbs are represented by a synset, which denotes a concept or a sense of a group of terms. These synsets are organized into taxonomic hierarchies via a variety of semantic relations. The semantic relations for nouns include Hyponym/Hypernym (is-a), Part Meronym/Part Holonym (part-of), Member Meronym/Member Holonym (member-of), Substance Meronym/Substance Holonym (substance-of) and so on. Hyponym/Hypernym (is-a) is the most common relations. Figure 1 illustrates a fragment of is-a hierarchy taxonomy in WordNet. In the taxonomy the deeper concept is more specific and the upper concept is more abstract.

Some algorithms have been proposed for semantic similarity metric based on WordNet in the past years. In this paper, we are only concerned about the similarity algorithm based on nouns and is-a relations of WordNet. Generally the typical algorithms based on WordNet can be grouped into two categories: edge-based similarity algorithms and information-based similarity algorithms. Next, we will give a brief review of these algorithms. Definitions of related concept in the following algorithms are as follows:

(1) len(c_1,c_2): the length of the shortest path from concept c_1 to concept c_2 in WordNet. *eg.*, len(boy,girl) is 4.

(2) lso(c_1,c_2): the most specific common subsumer of c_1 and c_2. *eg.*, lso(boy,girl) is person.

(3) depth(c): the length of the path to concept c from the global root entity. depth(root) is set to 1. *eg.*, depth (boy) is 7.

(4) deep_max: the max depth of the taxonomy. In Figure 1 deep_max is 8.

(5) hypo(c): the number of hyponyms for a given concept c. *eg.*, hypo (person) is 4.

(6) node_max: the maximum number of concepts that exist in the taxonomy of WordNet. In Figure 1 node_max is 25.

(7) sim (c_1,c_2): semantic similarity between concept c_1 and concept c_2.

2.1. Path-based Similarity Algorithms

Path-based similarity algorithms proceed from the position of each concept in the taxonomy to obtain semantic similarity. It assumes that the similarity between two concepts was the function of path length and depth. For two nodes that present concepts in the taxonomy, the smaller geometric distance they have, the more similar they are.

In a paper on translating English verbs into Mandarin Chinese, Wu and Palmer presented a scaled measure between a pair of concepts c_1 and c_2, which was defined as [12]:

$$sim_{W\&P}(c_1,c_2) = \frac{2*depth(lso(c_1,c_2))}{len(c_1,c_2)+2*depth(lso((c_1,c_2))} \tag{1}$$

It is noticed that the similarity between two concepts (c_1, c_2) is affected by two factors. One is the length of the shortest path from concept c_1 to concept c_2 (*ie.*, length(c_1,c_2)). The other is the specific common subsumer (*ie.*, lso(c_1, c_2)). Sim$_{W\&P}$ (c_1,c_2) is inversely proportional to length (c_1, c_2) and proportional to depth (lso(c_1, c_2)). If c_1 and c_2 are the same node in the taxonomy, len(c_1,c_2) is 0 and sim$_{W\&P}$ (c_1,c_2) get the max value 1, else 0< sim$_{W\&P}$ (c_1,c_2) < 1. Therefore, the values of sim$_{W\&P}$ (c_1,c_2) are in (0, 1].

Leakcock and Chodorow took the maximum depth of taxonomy into account and proposed the following measure [13]:

$$sim_{L\&C}(c_1,c_2) = -\log\frac{len(c_1,c_2)}{2*deep\_\max} \tag{2}$$

For a specific version of WordNet, deep_max is a fixed value, therefore sim$_{L\&C}$(c_1,c_2) is depended on the shortest path from c_1 to c_2 (*ie.*, length(c_1,c_2)).. If c_1 and c_2 are the same node in the taxonomy, len(c_1,c_2) is 0. In practice, we may add 1 to both len(c_1,c_2) and 2*deep_max to avoid log(0). Thus the values of $sim_{L\&C}$ (c_1, c_2) are range from 0 to log (2*deep_max+1).

Li, *et al.*, [14] combined the shortest path and the depth of concepts in a non-linear function, expressed by:

$$sim_{Li}(c_1,c_2)=e^{-\alpha*len(c_1,c_2)}\frac{e^{\beta*depth(lso(c_1,c_2))}-e^{-\beta*depth(lso(c_1,c_2))}}{e^{\beta*depth(lso(c_1,c_2))}+e^{-\beta*depth(lso(c_1,c_2))}} \quad (3)$$

Where α (α>0) and β (β>0) are parameters and used to adjust the contribution of shortest path length (*ie.*, length (c_1, c_2)) and depth (*ie.*, depth (lso (c_1, c_2)) respectively, which need to be adapted manually for good performance. In our experiment the same as in literature [14]'s, α is set to 0.2 and β is set to 0.6. It is noted that sim_{Li} (c_1, c_2) will increasing with respect to depth (lso (c_1, c_2)) and decreasing with len (c_1, c_2). The values of sim_{Li} (c_1, c_2) are between 0 and 1.

2.2. Information Content based Similarity Algorithms

Information content based algorithms do not consider the position, structure of each concept node in the taxonomy. It is based on the assumption that each concept contains a certain amount of information. The similarity is related to the information in common of concepts. The more common information two concepts share, the more similar the concepts are.

In 1995 Resnik proposed information content based similarity measure [15]. It was based on the assumption that for two given concepts, similarity is depended on the information content that subsumes them in the taxonomy. Hereafter Lin [16], Jiang [17], took the IC of compared concepts into account and proposed another two algorithms respectively. But their usage of IC was different.

Resnik assumed that for a concept c,

$$IC=-\log p(c) \quad (4)$$

Where p(c) is the probability of encountering and instance of concept c.

Probability of a concept was estimated as follows:

$$p(c)=\frac{freq(c)}{N} \quad (5)$$

Where N is the total number of nouns, and freq(c) is the frequency of instance of concept c occurring in the taxonomy.

When computing freq(c), each noun or any of its taxonomical hyponyms that occurred in the given corpora was included.

$$Freq(c)=\sum_{w\in W(c)}count(w) \quad (6)$$

Where *W(c)* is the set of words subsumed by concept c.

For two given concepts, similarity is depended on the information content that subsumes them in the taxonomy.

$$sim_{\text{Res}nik}(c_1,c_2) = -\log p(lso(c_1,c_2)) = IC(lso(c_1,c_2)) \tag{7}$$

Lin proposed another algorithm to measure similarity [16]. It assumed that the similarity between c_1 and c_2 was measured by the ratio between the amount of information needed to state the commonality of c_1 and c_2 and the information needed to fully describe what c_1 and c_2 were.

$$sim_{Lin}(c_1,c_2) = \frac{2 * IC(lso(c_1,c_2)))}{IC\ (c_1) + IC\ (c_2)} \tag{8}$$

Jiang calculated semantic distance to obtain semantic similarity [17]. Semantic similarity is the opposite of the distance.

$$dis_{Jiang}(c_1,c_2) = (IC(c_1) + IC(c_2)) - 2IC(lso(c_1,c_2)) \tag{9}$$

It is noted that the IC value of each concept plays an important role in information content based similarity algorithms. It is an important dimension in assessing the similarity of two concepts or two words and provides an estimation of its abstract or specialty. Generally speaking, there are two methods to obtain IC. One is Corpora-dependent IC metric. Corpora-dependent IC metric obtains IC through statistical analysis of corpora. The other is Corpora-independent IC metric. Recent years the latter has drawn great concern. One commonly used IC model was proposed by Nuno. The model use WordNet as a statistical resource to compute the probability of occurrence of concepts. It is based on the assumption that in WordNet IC value of a concept is regarded as the function of the hyponyms it has. Concepts with more hyponyms express less information than the concepts with less ones. It is defined as [18]:

$$IC(c) = 1 - \frac{\log(hypo(c)+1)}{\log(node\_\max)} \tag{10}$$

Obviously two concepts with the same number of hyponyms have the same IC value.

To overcome this problem, Meng took the topology structure of each concept into account and proposed another IC model, which can distinguish different concepts effectively and get more accurate IC value. It was defined as [19]:

$$IC(c) = \frac{\log(depth(c))}{\log(deep\_\max)} * (1 - \frac{\log(\sum_{a \in hypo(c)} \frac{1}{depth(a)} + 1)}{\log(node\_\max)}) \tag{11}$$

Where for a given concept c, a is a concept of the taxonomy, which satisfies $a \in hypo(c)$. If c is root, deep (root) is 1 and log (deep(c)) is 0. If c is a leaf, hypo(c) is 0. Then,

$$\sum_{a \in hypo(c)} \frac{1}{depth(a)} = 0 \tag{12}$$

And

$$IC(c)=\frac{\log(depth(c))}{\log(deep\_\max)}) \qquad (13)$$

3. A New Algorithm for Semantic Similarity Metric

The algorithm mentioned in Section 2 are all simple, but they can not distinguished different pairs effectively. In path based algorithms, every edge is equal. In Figure 1 take pairs (vehicle, mail) and pairs (bus, train) as an example, we find that len(vehicle, mail) is equal to len(bus, train). Thus the two pairs will have the same similarity values with with Leakcock and Chodorow's method. A fact must be noticed that pairs(bus,train) is deeper in the taxonomy and more concrete than pairs(mail,vehicle), thus pairs(bus,train) should be more similar than pairs(vehicle,mail).

Another example are pairs(mail, bicycle) and pairs(wheeled vehicle, bus), we find that:

(1) Len(mail, bicycle) is equal to len(wheeled vehicle, bus).

(2) Both lso (mail, bicycle) and lso (wheeled vehicle, bus) are conveyance.

This fact make the two pairs will have the same similarity value with Wu&Palmer's method, Li's method and Resnik's method respectively. This is not reasonable.

Next, let's look at Lin's Method and Jiang's method.

Accordng to formula (10) IC (mail) = IC (bicycle) = IC (school bus) = 1. Pairs (mail, bicycle) and pairs (bicycle, school bus) will have the same similarity value. Therefore different concepts pairs could not be distinguished effectively. Despite of we will get different similarity values according to formula (11), the result is not very close to human's judgment. There is still room for improvement.

Here, a new algorithm is proposed. It assumes that the similarity between concept C_1 and C_2 is a function of path length and local density, which is defined as:

$$sim(c_1,c_2)=(\frac{2*IC(lso)}{IC(c_1)+IC(C_2)})^{(k*(1-e^{-len(c_1,c_2)}))} \qquad (14)$$

where k is a parameter, which can be adapted manually. Experiments show that when k is 1.6, the algorithm will get the best performance.

In the new algorithm, both path length between two concepts and their information contents have been taken into account. In Figure 1, in spite of len(mail, bicycle) is equal to len(wheeled vehicle, bus), and the two pairs have the same most specific subsumer, but they convey different information content. Thus, according to formula (14), len(mail, bicycle) and len(wheeled vehicle, bus) will have different semantic similarity values, which can effectively avoid the problems discussed above.

Besides this, from formula (14), we can see that,

Firstly, $e^{-len(c_1,c_2)}$ is a nonlinear, decreasing function with respect to $len(c_1,c_2)$.

Secondly, $Sim(c_1,c_2)$ is inversely proportional to $len(c_1,c_2)$. If two concepts have the same sense, $len(c_1,c_2)$ is 0, and $sim(c_1,c_2)$ is 1.

Finally, sim(c_1,c_2) are in (0, 1].

In next section, we will analyze our new method from different perspectives.

4. Evaluation

In this section, we compare the six chosen methods listed in Section 2 with our new method by how well they reflect human's judgments.

4.1. Data Set

For evaluating the performance of our new algorithm, a dataset is necessary. In the experiment, we adopt the dataset provided by Rubenstein and Goodenough (1965) [20], which are commonly used. In their study, 51 undergraduate subjects were given 65 pairs of words, which ranged from "highly synonymous" to "semantically unrelated". Subjects were asked to rate them on the scale of 0.0 to 4.0. The average rating for each pair reflects a good estimate of how similar the two words are.

4.2. Words Similarity Calculating Method

Because either or both of the words have more than one sense in WordNet, we need to compute the semantic similarity matrix as Table 1.

Table 1. Semantic Similarity Matrix

Word Pairs		*Word$_2$*			
		c_{21}	c_{22}	……	c_{2j}
Word$_1$	c_{11}	$sim_{11\text{-}21}$	$sim_{11\text{-}22}$	……	$sim_{11\text{-}2j}$
	c_{12}	$Sim_{12\text{-}21}$	$sim_{12\text{-}22}$	……	$sim_{12\text{-}2j}$
	……	……	……	……	……
	c_{1i}	$sim_{1i\text{-}21}$	$sim_{1i\text{-}22}$	……	$sim_{1i\text{-}2j}$

Where C_{1i} is the sense of word1, and C_{2j} is the sense of word2. In the result, we took the most similarity pair of sense:

$$sim(Word_1, Word_2) = \max_{(i,j)}[sim(c_{1i}, c_{2j})]$$

For each of seven implemented algorithm, we compute similarity scores for the human-rated pairs.

4.3. Results Analysis

Before our analysis, we first compute semantic similarity between pairs of words with formula (1) ~ (3), (7) ~ (9) and our new algorithm, and draw the distributed graph in Figure 2. For the convenience of expression and comparison, we normalized the values in [0, 1]. The IC value is obtained according to formula (10), (11) respectively.

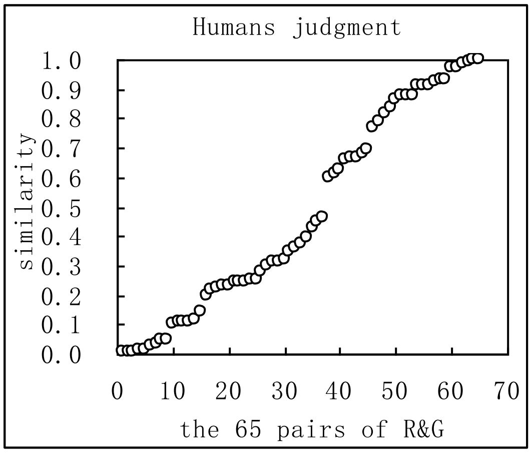

Figure 2(1). Human's Similarity

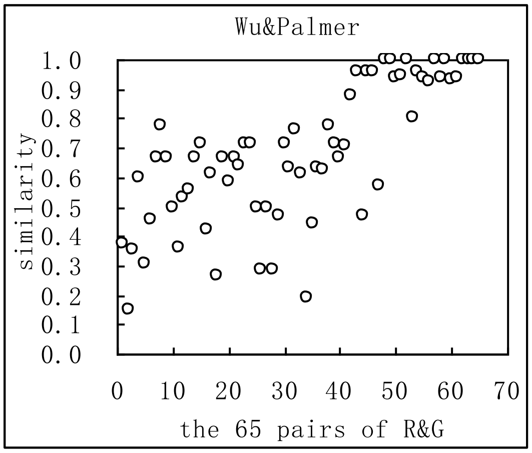

Figure 2(2). Wu&Palmer's Similarity

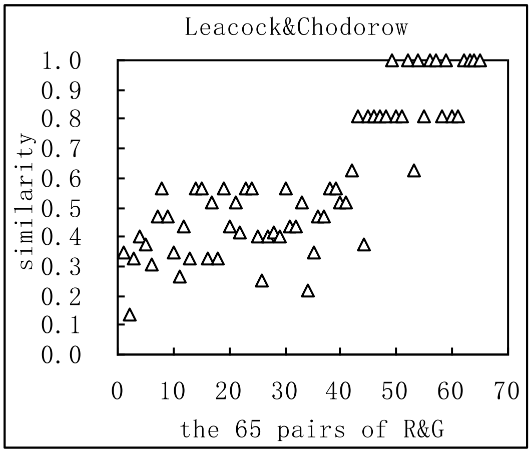

Figure 2(3). Leacock & Chodorow's Similarity

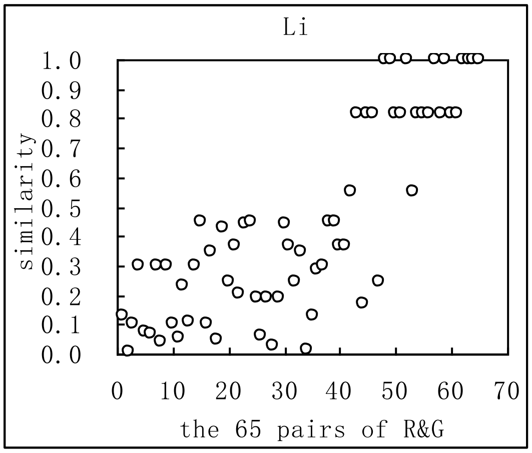

Figure 2(4). Li's Similarity

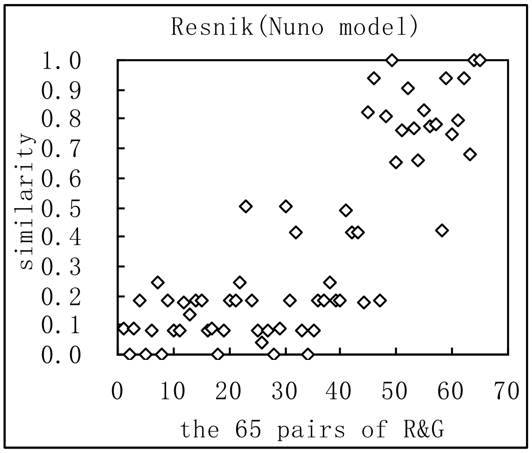

Figure 2(5). Resnik's Similarity (Nuno)

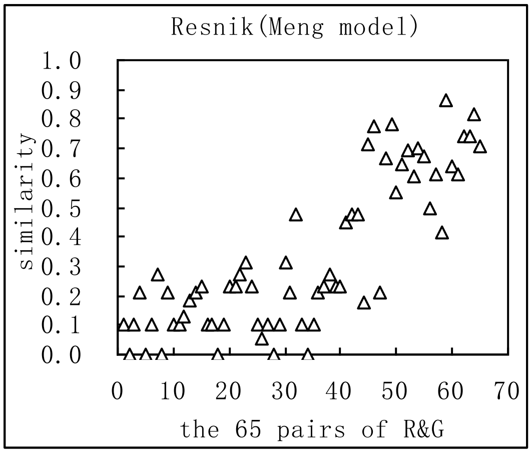

Figure 2(6). Resnik's Similarity (Meng)

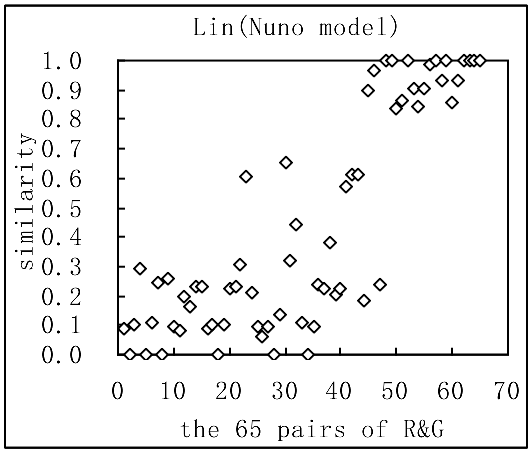

Figure 2(7). Lin’s Similarity (Nuno)

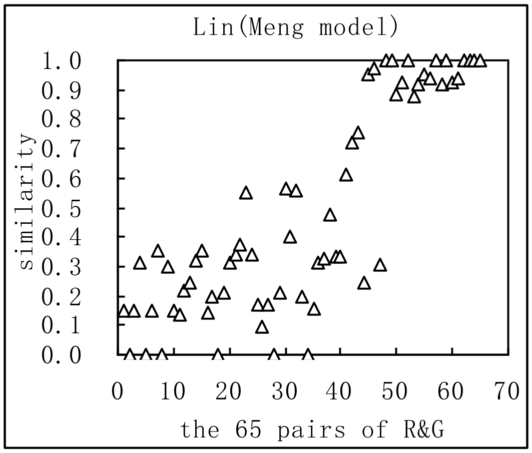

Figure 2(8). Lin’s Similarity (Meng)

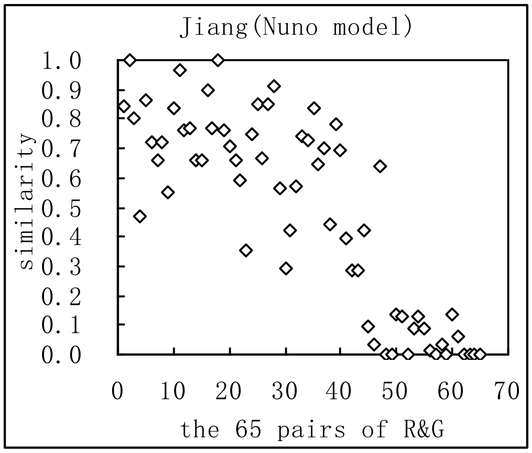

Figure 2(9). Jiang’s Similarity (Nuno)

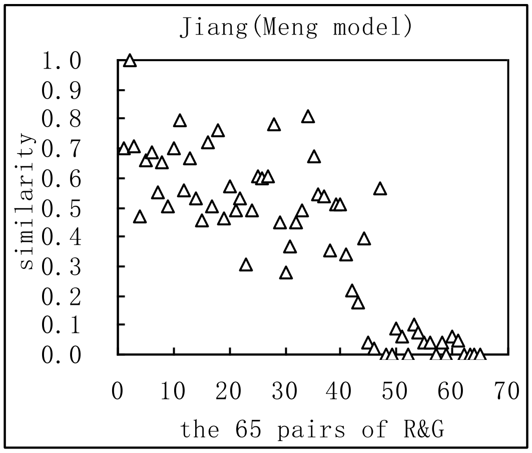

Figure 2(10). Jiang’s Similarity (Meng)

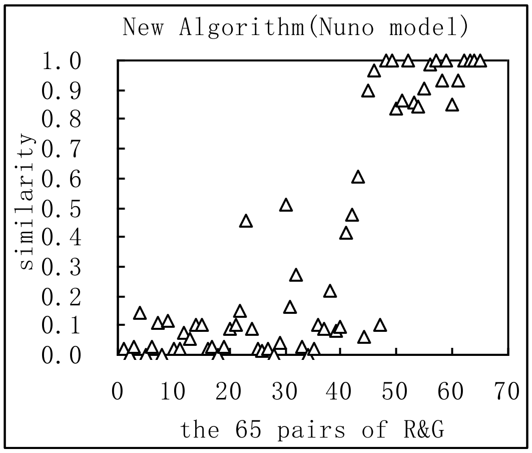

Figure 2(11). Our’s Similarity (Nuno)

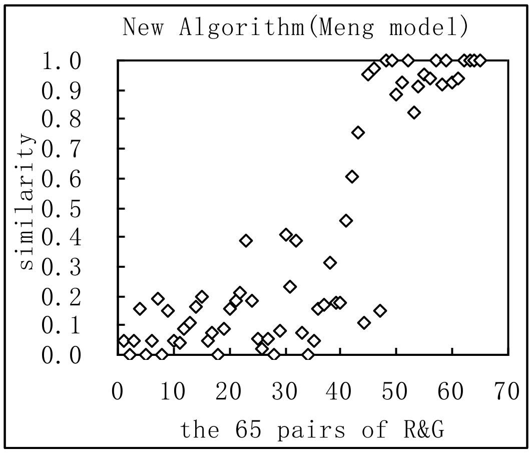

Figure 2(12). Our’s Similarity (Meng)

In accordance with previous research, we compare the six chosen algorithms mentioned in Section 2 with our new algorithm by calculating the coefficients of correlation with human judgments of semantic similarity.

Table 2 presents the chosen algorithms and their correlation coefficient.

Figure 3 shows the compared results of our proposed algorithm with other six methods.

Table 2. Correlation Coefficients with Different Algorithms

Similarity Algorithm			Correlation Coefficient (R&G)
Path baesd	Wu & Palmer (W&P)		0.7767
	Leacock & Chodorow (L&C)		0.8535
	Li		0.8559
Information content based	Resnik	Nuno model	0.8400
		Meng model	0.8487
	Lin	Nuno model	0.8643
		Meng model	0.8728
	Jiang	Nuno model	-0.8569
		Meng model	-0.8670
New algorithm		Nuno model	0.8697
		Meng model	0.8820

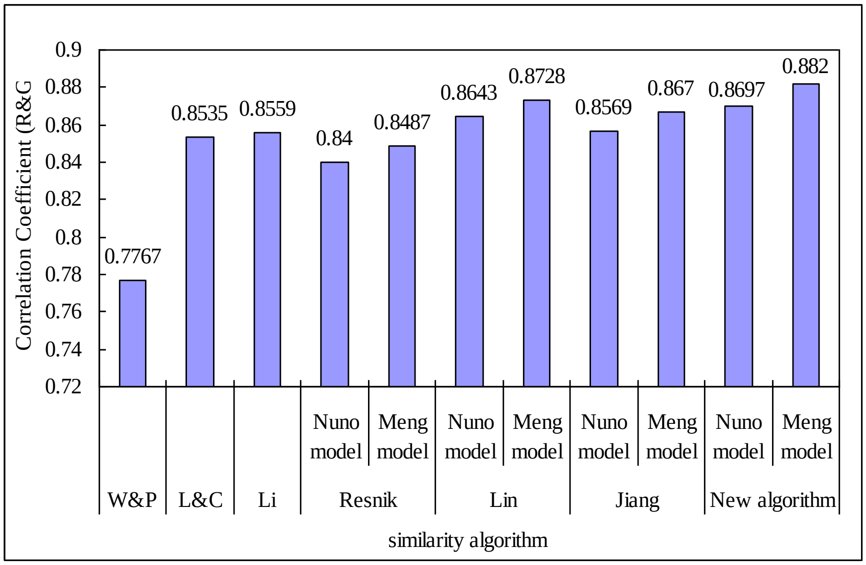

Figure 3. Compare Our Proposed Algorithm with Others

From Table 2 and Figure 3, we can see that,

Firstly, the correlations coefficient values with Meng model are more superior to the ones with Nuno model in Resnik's algorithm (Meng model: 0.8487, Nuno model: 0.8400), Lin's algorithm (Meng model: 0.8728, Nuno model: 0.8643) and Jiang's algorithm (Meng model: -0.8670, Nuno model: -0.8569) and our new algorithm (Meng model: 0.8820, Nuno model: 0.8697) respectively.

Secondly, the correlations coefficient in our new algorithm is higher than any other path based similarity algorithms and information content based similarity algorithms whether with Nuno model (0.8697) or Meng model (0.8820).

Finally, our new algorithm has achieved the best performance with Meng model (0.8820).

5. Conclusion and Future Work

This paper presents a novel algorithm of semantic similarity metric of word pairs based on WordNet. Different from previous work, in the new algorithm both path length and IC value have been taken into considerate. We evaluate the performance of our new algorithm on the data set of Rubenstein and Goodenough (1965), which is traditional and widely used. The distributed graphs of 65 word pairs' similarity values with different algorithms are illustrated. Experiments show that the correlation with human judgment is 0.8820, which demonstrated that the proposed algorithm significantly outperformed traditional similarity algorithms. In future work, we will attempt to use this model in real world applications such as word sense disambiguation, information retrieval, ontology construction and so on.

References

[1] M. R. Quilian, "Semantic memory", In M. Minsky, Ed., Semantic Information Processing, MIT Press, Cambridge, MA, **(1968)**.

[2] S. Patwardhan, S. Banerjee and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation", Proceedings of 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, **(2003)** February 16-22; Mexico City, Mexico.

[3] J. Atkinson, A. Ferreira and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts", Knowl.-Based Syst., vol. 22, no. 7, **(2009)**.

[4] M. Stevenson and M. A. Greenwood, "A semantic approach to IE pattern induction", Proceedings. of 43rd Annual Meeting on Association for Computational Linguistics, **(2005)** June 25-30; Ann Arbor, Michigan, USA.

[5] D. Sánchez, D. Isern and M. Millán, "Content annotation for the Semantic Web: an automatic web-based approach", Knowl. Inf. Syst., vol. 27, no. 3, **(2011)**.

[6] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using ngram co-occurrence statistics", Proceedings of Human Language Technology Conference, **(2003)** May 27-June 1; Canada, Edmonton.

[7] A. G. Tapeh and M. Rahgozar, "A knowledge-based question answering system for B2C eCommerce", Knowl.-Based Syst., vol. 21, no. 8, **(2008)**.

[8] Y. Blanco-Fernández, J. J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, M. López- Nores, J. García-Duque, A. Fernández-Vilas, R. P. Díaz-Redondo and J. Bermejo-Muñoz, "A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems", Knowl.-Based Syst., vol. 21, no. 4, **(2008)**.

[9] H. Kozima, "Computing Lexical Cohesion as a Tool for Text Analysis", doctoral thesis, Computer Science and Information Math ,Graduate School of Electro-Comm., Univ. of Electro-Comm., **(1994)**.

[10] G. Grefenstette, "Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches", Workshop on acquisition of lexical knowledge from text, Columbus, OH, **(1993)**.

[11] C. Fellbaum, Ed., "WordNet: An Electronic Lexical Database", MIT Press, Cambridge, USA, **(1998)**.

[12] Z. Wu and M. Palmer, “Verb semantics and lexical selection”, Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, **(1994)** June 27-30; Las Cruces, New Mexico.

[13] C. Leacock and M. Chodorow, “Combining Local Context and WordNet Similarity for Word Sense Identification”, WordNet: An Electronic Lexical Database, MIT Press, **(1998)**, pp. 265-283.

[14] Y. Li, A. B. Zuhair and D. McLean, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”, IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, **(2003)**.

[15] P. Resnik, “Using information content to evaluate semantic similarity”, Proceedings of the 14th International Joint Conference on Artificial Intelligence, **(1995)** August 20-25; Montr éal Qu ébec, Canada.

[16] D. Lin, “An information-theoretic definition of similarity”, Proceedings of the 15th International Conference on Machine Learning, **(1998)** July 24-27; Madison, Wisconsin, USA.

[17] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy”, Proceedings of International Conference on Research in Computational Linguistics, **(1997)** August 22-24; Taipei, Taiwan.

[18] N. Seco, T. Veale and J. Hayes, “An intrinsic information content metric for semantic similarity in WordNet”, Proceedings of the16th European Conference on Artificial Intelligence, **(2004)** August 22-27; Valencia, Spain.

[19] L. Meng, J. Gu and Z. Zhou, “A New Model of Information Content Based on Concept's Topology for measuring Semantic Similarity in WordNet”, International Journal of Grid and Distributed Computing, vol. 5, no. 3, **(2012)**.

[20] H. Rubenstein and J. B. Goodenough, “Contextual correlates of synonymy”, Communications of the ACM, vol. 8, no. 10, **(1965)**.

Authors

Lingling Meng is a PhD Candidate of Computer Science and Technology Department and a teacher of Department of Educational Information Technology in East China Normal University. Her research interests include intelligent information retrieval, ontology construction and knowledge engineering.

Runqing Huang has a PhD from Shanghai Jiao Tong University. He works in Shanghai Municipal People's Government, P. R. China. His present research interests include modeling strategic decisions, economic analysis, electronic government and Logistics.

Prof. Junzhong Gu is Supervisor of PhD Candidates, full professor of East China Normal University, head of Institute of Computer Applications and director of Lab, Director of Multimedia Information Technology (MMIT). His research interests include information retrieval, knowledge engineering, context aware computing, and data mining.